DOCUMENT RESUME

ED 137 386 TM 006 192

AUTHOR Garrison, Wayne M.; And Others
TITLE The Use of the Rasch Measurement Model in Developmental Analyses of Children's Self-Concept.
SPONS AGENCY National Inst. of Mental Health (DHEW), Rockville, Md.; Purdue Research Foundation, Lafayette, Ind.
PUB DATE [Apr 77]
GRANT 0074-56-12675; R01-MH-19384-02
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Affective Behavior; Analysis of Variance; *Behavior Change; Elementary Education; *Elementary School Students; Females; *Item Analysis; Longitudinal Studies; Males; *Mathematical Models; *Self Concept; Self Concept Tests
IDENTIFIERS Piers Harris Self Concept Scale; *Rasch Model

ABSTRACT

The utility of the Rasch logistic measurement model in longitudinal analyses of children's self-concept was investigated. Calibration of an 80-item self-concept measure from the responses of 1,000 elementary school children resulted in the identification of 25 items that fit the model. Repeated measures analyses of variance used to assess the effects of an experimental intervention program on changes in individuals' self-concept tended to support the usefulness of the measure in reducing the number of test items needed to measure the latent variable in question. Some evidence on the claim for item subtest equivalence also was obtained. (Author/RC)

The Use of the Rasch Measurement Model in Developmental Analyses of Children's Self-Concept[1, 2]

Wayne M. Garrison
National Technical Institute for the Deaf

Allan S. Cohen and Kathryn W. Linden
Purdue University

The assessment of change in affective behaviors has become an important concern of educators, particularly in instances where evidence is sought regarding the effects of planned interventions on affective outcomes. In areas of research on self-concept in particular, several studies have attempted to construct a developmental picture of this aspect of human behavior (e.g., Abbele, 1967; Carpenter & Busse, 1969; Stanwyck & Felker, 1971) for the purpose of understanding further the nature of the growth or change process. Whether or not these studies adequately addressed the issue of age changes in self-concept remains an open question due to the use of cross-sectional research methodology. However, longitudinal research also has been conducted in this regard (Felker, 1972, 1976; O'Malley and Bachman, 1976; Stanwyck, 1972).

---

[1]This paper is based in part on the first author's doctoral dissertation at Purdue University.

[2]This research was supported by an NIMH grant to Donald W. Felker (RO1-MH-19384-02) and by a grant from the Purdue Research Foundation (0074-56-12675) to the second author.

Unfortunately, a review of existing research literature on self-concept reveals inconsistencies and wide-ranging differences in the findings reported. It may be that comparisons among studies have suffered most because of differences in the instruments used to measure the relevant behaviors, lack of definitional clarity in the concepts studied, as well as in the sampling procedures employed, and even procedural differences in testing conditions.

From a psychometric perspective, still another problem arises in the interpretation of research findings derived from the use of repeated measurements from a specified population of test-takers. Addressing this problem, Anderson (1976) posed the question as to whether growth or change scores on a particular affective measure could be interpreted as reflecting behavioral changes on the variable(s) of interest among the individuals tested, or whether the test items themselves undergo some change over time. That is, Anderson suggested that the use of traditional measurement models may confound behavioral changes among persons with changes over time in the item characteristics themselves. What would be useful, then, in order to measure behavioral change on some psychological construct would be a set of items whose psychometric properties remain invariant over time.

Coupled with the measurement concerns expressed above is the problem of generalizability of research findings derived from an experimentally accessible population to some

theoretically larger target population. Because such generalizations require thorough knowledge of the characteristics of both the sampled and intended populations (information that may be either unavailable or unknown to a researcher), generalizability remains a major obstacle in behavioral research. On this issue, Rasch (1960) noted the strong interdependence between statistical tools and characteristics of the particular sample of persons selected for study. When traditional models of measurement are used, Rasch demonstrated that the psychometric properties of tests are not specific to the instruments themselves and may vary markedly with the sample studied. Thus, an individual's score on a test is largely dependent for its meaning upon a particular set of items and a particular sample of test-takers.

During the last decade, the topic of latent trait models has received the attention of measurement specialists as a means of improving educational assessment practices. The particular model advanced by Rasch (1960) has been described as providing individual measurements of behavior that are independent of either the sample of persons from whom the measurements were obtained or the particular set of items used to measure a given behavior. Moreover, it has been claimed that, if an instrument can be demonstrated as fitting the model, any subset of calibrated items will provide comparable measures of the behavior in question. Thus, Rasch has

suggested that an instrument possessing the general characteristics of his measurement model would become analogous to a yardstick used to measure the length of physical objects.

The purpose of this study was to examine the usefulness of the Rasch logistic measurement model for longitudinal research on change in affective behaviors of children. Specifically, evidence was sought regarding the degree to which the Rasch model claims were substantiated in the measurement of affective behavioral outcomes. In testing the claims of the model, special attention was given to the issue of item subset equivalence.

METHODOLOGY

Examinee Population

The primary sample for this study consisted of 1,927 elementary-school children for whom measures of self-concept had been obtained during September, 1972, May, 1973, and September, 1973, as part of a longitudinal study of children's self-concept development (Felker, 1972). Testing was conducted under classroom conditions using four schools in northwestern Indiana.

Instrumentation

The Piers-Harris Self-Concept Scale (Piers & Harris, 1964) was used in the present study as the measure of self-concept. In its original form, the scale includes 80 declarative statements, originally developed from Jersild's (1952) categories, and requires the examinee to respond either yes or no on the

basis of whether or not each statement is congruent with the examinee's perception of him or herself. The authors reported that the instrument was developed as a measure of general self-concept. However, a recent review of research on the instrument (Shavelson, Hubner & Stanton, 1976) suggests the presence of several different dimensions, with general self-concept (total scores on the scale) possibly reflecting a relatively enduring characteristic of the individual test-taker.

Design

The test responses obtained during September, 1972, from a random sample (without replacement) of 1,000 subjects drawn from the total examinee population were used to calibrate the Piers-Harris Scale. Because calibration of an instrument rarely is accomplished in a single computer run, a series of analyses was required in order to produce a final set of items whose properties satisfied the assumptions underlying the application of the Rasch measurement model. The CALFIT computer algorithm (Wright & Mead, 1975) was used to calibrate the scale and, hence, to estimate the Rasch person and item parameters explicit in the measurement model.

Following the sequence of test calibrations noted above, a final set of 25 items (from an original pool of 80) that fit the Rasch formulation was obtained. To test the model's claim that any subset of items from a calibrated pool of items may be used to provide comparable measures of the construct in question, subsets of 16 items each were drawn randomly (without replacement) from the 25 calibrated items. The

determination of the number of items to be included in these subtests (k = 16) was based upon the results of a study conducted by Garrison (1976), indicating that this number of items was necessary to establish an average stability coefficient of .65 between testing times.

In order to examine the effectiveness of an intervention program (Felker, 1972) on the development of individuals' self-concept, the test data were analyzed using a repeated measures analysis of variance design (Winer, 1971). Analyses were performed separately for each sex and consisted of two between (or crossed) factors (experimental vs. control treatment groups and grade level of respondents) and one within (or nested) factor (time of testing). Thus, the dependent measures for the analyses consisted of total scores obtained by examinees on each of the following instrument forms: (1) the 80-item original Piers-Harris Scale, (2) a 25-item instrument composed of items fitting the measurement model and (3) fifteen (15) 16-item tests composed of items drawn randomly from the set of 25 calibrated items. The results of each of the repeated measures analyses of variance were compared to determine whether self-concept changes over time were manifested consistently for each of the differing test formats studied.

RESULTS

Using the response data obtained from 924 male and female pupils in grades 2 - 5 over a one-year period, repeated measures analyses of variance were used to examine the consistency of self-concept changes manifested over time when the test stimuli were manipulated deliberately. In order to avoid confounding of sex differences in self-concept development with the more meaningful differences under investigation in this study, the data obtained from males and females were analyzed separately. Table 1 summarizes the results of analyses using scales composed of 80, 25, and 16-items for the male examinees. Except for the mean square error terms, the cell entries represent F-ratios.

Insert Table 1 about here

The analyses performed on both the 80-item and 25-item sets utilized identical instruments, respectively, over three different times of measurement. However, it must be noted that each of the 16-item sets differed from one time of measurement to another. The variation among the 16-item scales was introduced in order to test the Rasch claim that "any" set of calibrated items may be used to measure a person's position along some latent continuum.

An examination of the data provided in Table 1 indicates that the significant effects observed for the 80-item self-concept instrument generally continued to be manifested with the use of 25 and 16 items. In particular, the main effect

due to the experimental treatment was consistent. Upon further examination, it was found also that, for each comparison of the treatment means, the experimental group scored significantly higher than did the control group of respondents. For all analyses, total test scores were expressed in the Rasch log metric.

The T (time of measurement) main effect also showed consistency among the analyses presented in Table 1 and, generally, reflected an upward movement from Time 1 to Time 3 in terms of mean self-concept. In the case of the treatment by time of measurement interaction (A x T), the findings were not as clearly interpretable. That is, the significant interaction observed from the 80-item data was manifested in only 50% of the analyses performed using fewer items.

The results of analyses of variance performed on the self-concept data for the female examinees are presented in Table 2.

Insert Table 2 about here

As was found for the results of analyses performed using the male respondents, both the A and T main effects tended to show consistency over the differing test formats. However, the A x B and A x B x T interactions observed for the 80-item data set consistently failed to appear in subsequent analyses using fewer items. While the reason for this finding was not

clear, it was speculated that the reduction in total test score variance for the 16-item subtests (attributable to test length and similar item discrimination indices) may have accounted for the "loss of power" in uncovering these subtle interactions.

A somewhat different pattern of results than those described above was obtained from an examination of the data contained in Table 3.

Insert Table 3 about here

Table 3 details the proportion of total variance attributable to each of the experimental effects reaching statistical significance for the test formats studied. While the differences in the proportions reported are probably too small to be theoretically important, it is interesting, nonetheless, to note the effects of reducing the total number of test stimuli on the resulting proportions.

For the male respondents, reducing the number of test items to 25 and 16, respectively, resulted in some gain in the proportion of variance attributable to the experimental intervention. However, whatever gain resulted for the male respondents was offset by a decrease in the same proportion of variance attributable to the experimental treatment among the female respondents. Thus, the effects of fitting a set of items to the Rasch logistic model resulted in some efficiency and

consistency of the changes manifested in males over time, while the same conclusion could be drawn only tenuously for the female population of respondents.

DISCUSSION AND CONCLUSIONS

The present study was conducted to examine the usefulness of the Rasch logistic measurement model in developmental studies of children's self-concept. A calibration analysis of the Piers-Harris Self-Concept Scale (Piers & Harris, 1964) resulted in the identification of a set of 25 test items that fit the assumptions of the model (Garrison, 1976). In fitting the model to the data, it was observed that a large proportion of the total number of items contained in the intact instrument were discarded. However, the reductions in number of test stimuli used to assess growth or change along an affective dimension did not appear to alter markedly the conclusions drawn from analyses of data collected using the larger item pool. Thus, if one is willing to accept slightly less control in an experimental context, it appears reasonable that Rasch calibration procedures may be useful in constructing tests which are unidimensional and considerably shorter in over-all length than those typically utilized in psychological research. More importantly, the unidimensional nature of Rasch calibrated tests (Rasch, 1960; Wright & Mead, 1975) may well serve to clarify the nature of the construct being measured.

Within the limits of statistical probability, then, the Rasch model was found to be useful in reducing the number of items required to measure growth or change along selected

psychological characteristics. Furthermore, as Anderson (1976) noted, if one is to accurately estimate change among individuals along an underlying construct, then it is necessary to develop instruments with meaningful units. However, the assertion that "any" set of calibrated items may be used to measure a person's position along some latent continuum represents a significant departure from the traditional approach to equating instruments. Yet, if it can be determined that item sets that have not been matched for difficulty provide the same information about persons as do instruments that have been equated using traditional procedures, then the classical test theory requirements for equating instruments (i.e., equal item difficulties, discriminations, means and variances) may be unnecessary.

In conclusion, research on the application of the Rasch measurement model has been limited primarily to the measurement of intelligence or achievement-related outcomes. Yet, an examination of the conditions specified by Wright and Mead (1975) for the use of the model suggests that it may have utility also in the measurement of affective behaviors. Future research on the validity of the model may seek to pursue evidence for the model's claims along longitudinal avenues. Whereas much attention has now been devoted to the robustness of the model's underlying assumptions, much more effort must be expended in establishing the model's utility within an experimental context.

REFERENCES

Abbele, F. M. The development of self-perception from ages 6 to 10 through the self-portrait. Studi e Ricerche di Psicologia, 1967, 10, 1-7.

Anderson, L. W. Using the Rasch model to measure change in academic self-concept resulting from instructional and curricular innovations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Carpenter, T. R., & Busse, T. V. Development of self concept in Negro and white welfare children. Child Development, 1969, 40, 935-939.

Felker, D. W. Self-concept and self-administered verbal rewards. Progress Report, February 1972, Purdue University, Grant RO1-MH-19384-02, National Institute of Mental Health.

Felker D. W. Self-concept and self-administered verbal rewards. Final Report, June 1976, Purdue University, Grant RO1-MH-19384-02, National Institute of Mental Health.

Garrison, W. M. The use of the Rasch measurement model in developmental self-concept research. Unpublished doctoral dissertation, Purdue University, 1976.

Jersild, A. T. In search of self. New York: Teachers College Press, 1952.

O'Malley, P. M., & Bachman, J. G. Self-esteem and educational attainment: A longitudinal analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Piers, E. V., & Harris, D. B. Age and other correlates of self-concept in children. Journal of Educational Psychology, 1964, 55, 91-95.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. Self-concept: Validation of construct interpretations. Review of Educational Research, 1976, 46, 407-441.

Stanwyck, D. J. Self-concept development: A longitudinal study. Unpublished doctoral dissertation, Purdue University, 1972.

Stanwyck, D. J., & Felker, D. W. Measuring the self-concept: A factor analytic study. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1971.

Winer, B. J. Statistical principles in experimental design (3rd Ed.). New York: McGraw-Hill, 1971.

Wright, B. D., & Mead, R. J. CALFIT: Sample-free item calibration with a Rasch measurement model. Research Memorandum 18, Statistical Laboratory, Department of Education, University of Chicago, March 1975.

Table 1

Summary of Analysis of Variance Results for Male Self-Concept Data (n = 494)

Source of Variation	df	Full Scale (k=80)	Calib. Scale (k=25)	Random I	Random II	Random III (k=16)	Random IV	Random V
Between Subjects								
Treatment Groups (A)	1	2.928*	11.443**	7.163**	10.043**	10.282**	7.137**	6.679**
Grade (B)	3	.209	.559	.252	.559	.232	1.458	.611
A x B	3	1.367	.939	.556	1.244	1.359	.458	.754
Error Mean Square	486	2.412	2.881	2.603	2.636	2.589	2.503	2.473
Within Subjects								
Time of Measurement (T)	2	14.419**	16.977**	3.171**	10.904**	13.687**	12.854**	13.805**
A x T	2	3.105**	2.397*	.947	3.175**	2.062	2.238*	1.856
B x T	6	.827	.552	.985	.600	.653	.507	1.937*
A x B x T	6	1.041	.970	1.371	.588	.839	1.505	1.962*
Error Mean Square	972	.349	.540	.572	.546	.557	.561	.550

Note: Except for error mean square terms, the cell entries represent F-ratios.

* $p < .10$

** $p < .05$

Table 2

Summary of Analysis of Variance Results for Female Self-Concept Data (n = 430)

Source of Variation	df	Full Scale (k=80)	Calib. Scale (k=25)	Random I	Random II	Random III (k=16)	Random IV	Random V
Between Subjects								
Treatment Groups (A)	1	6.059**	4.540**	3.131*	3.221*	3.288*	.986	3.592*
Grade (B)	3	.569	.922	.888	.854	.674	.619	1.131
A x B	3	2.311*	1.287	1.349	1.343	1.091	1.586	.784
Error Mean Square	422	2.635	3.301	2.797	2.934	2.905	2.859	2.745
Within Subjects								
Time of Measurement (T)	2	14.990**	17.565**	6.851**	20.749**	16.694**	10.120**	19.946**
A x T	2	.586	2.554*	.126	3.033**	1.370	.538	.047
B x T	6	.741	2.147**	2.165**	1.596	2.028*	2.534**	1.726
A x B x T	6	2.065*	1.088	.516	.774	1.075	1.175	.698
Error Mean Square	844	.339	.516	.547	.566	.556	.553	.530

Note: Except for error mean square terms, the cell entries represent F-ratios.

* $p < .10$

** $p < .05$

Table 3

Proportion of Total Variance
Attributable to Significant Effects for Differing Test Formats

Sex of Respondents	Significant Effects Observed	Number of Items 80	25	16*
MALES	A	.0046	.0165	.0115
	T	.0065	.0092	.0066
	AT	.0014	.0013	.0012
	Error	.9776	.9641	.9704
FEMALES	A	.0110	.0079	.0047
	AB	.0126	.0067	.0061
	T	.0070	.0096	.0095
	ABT	.0029	.0018	.0016
	Error	.9622	.9643	.9694

* Proportions appearing in this column represent averages based upon the analyses of variance of the 15 sets of 16 items randomly drawn from the calibrated set of 25 items.